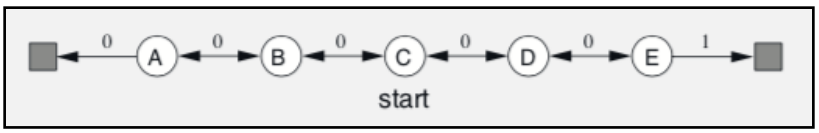
RL Assignment 3

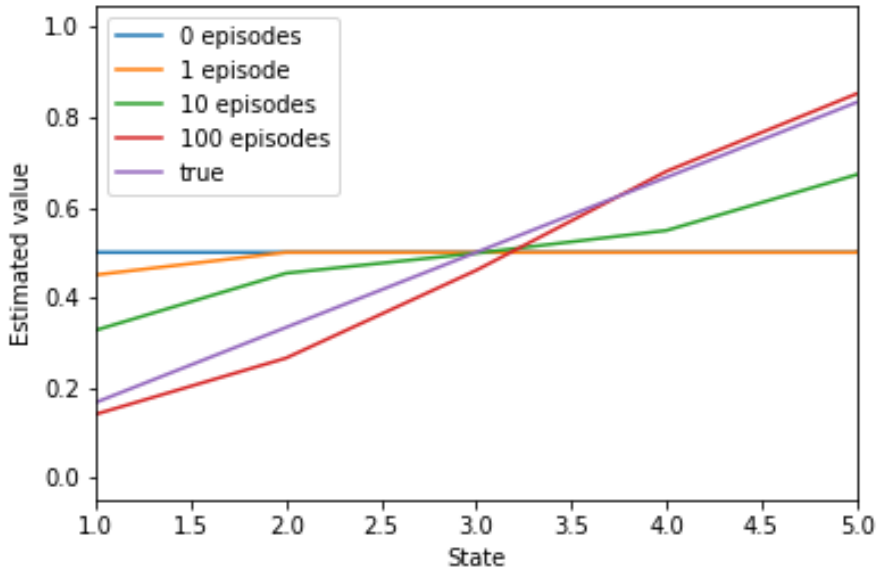
Srivatsava Kesanupalli
MT18054

Q6. Random walk problem to compare TD(0) and constant-α MC

In the problem given the following Markov Reward Process

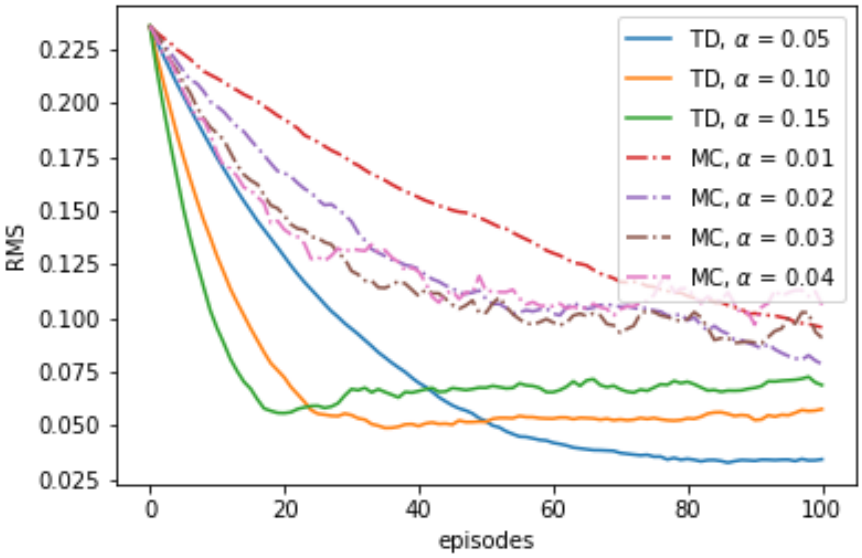


The agent starts at state C and moves to either of the terminal states and this concludes an episode. Initially the state values are set to 0.5 for every state. The reward on moving to the next state is 0 and if the agent moves to the left terminal state, the reward is 0 and the episode terminates. If the agent terminates on the right terminal state, the reward is +1 and the episode terminates. In this way, the state values are plotted for each of the following episodes 0, 1, 10 and 100 and the graph plots can be seen below



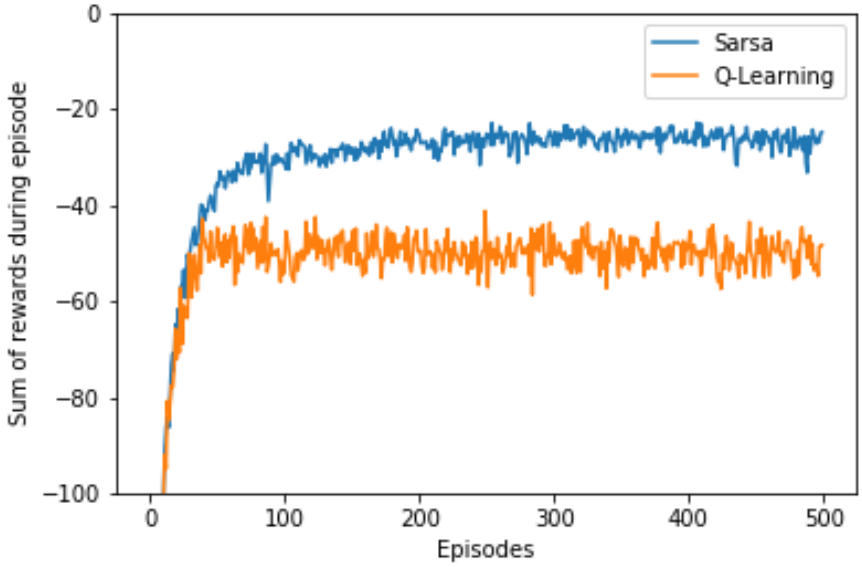
As the agent completes 100 episodes, the state values move close to that of the true values. After a certain number of episodes, the algorithm converges to true values. This demonstrates the convergence abilities of the Temporal Difference algorithm.

We know, repeat the same process with multiple α values for both Monte Carlo and Temporal Difference algorithms. The root mean squared error between true values and current values of the states are plotted for r.m.s values averaged over 100 runs in each episode. The procedure is repeated for 100 episodes and the way error values varied can be seen in the below plot. Differences for α-values is explained in the Exercise 6.4, 6.5



Q6. Cliff walking problem to compare Q-Learning and Sarsa

Differences with Q-Learning and Sarsa are explained in the Exercise 6.12
We can see Q-Learning vs. Sarsa plot for sum of rewards during episode vs. episode in the below graph plot



Because of ε-greedy action selection, the Q-Learning algorithm learns the values for the optimal policy, which travels right and eventually falls of the cliff, resulting in a reward of -100. Sarsa takes this into account and learns the safer path resulting in less sum of rewards. The above plot clearly shows the difference